Automatic Transformation of a Video Using Multimodal Information for an Engaging Exploration Experience

applied sciences

MDPI

Article

Automatic Transformation of a Video Using Multimodal Information for an Engaging Exploration Experience

Fahim A. Salim [1], Fasih Haider [2,*], Saturnino Luz [2] and Owen Conlan [1]

[1] ADAPT Centre, Trinity College Dublin, D02 PN40 Dublin, Ireland; salimf@tcd.ie (F.A.S.); Owen.Conlan@scss.tcd.ie (O.C.)

[2] Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, EH16 4UX, UK; S.luz@ed.ac.uk

* Correspondence: Fasih.Haider@ed.ac.uk

Received: 21 March 2020; Accepted: 21 April 2020; Published: 27 April 2020

Abstract: Exploring the content of a video is typically inefficient due to the linear streamed nature of its media and the lack of interactivity. While different approaches have been proposed for enhancing the exploration experience of video content, the general view of video content has remained basically the same, that is, a continuous stream of images. It is our contention that such a conservative view on video limits its potential value as a content source. This paper presents An Alternative Representation of Video via feature Extraction (RAAVE), a novel approach to transform videos from a linear stream of content into an adaptive interactive multimedia document and thereby enhance the exploration potential of video content by providing a more engaging user experience. We explore the idea of viewing video as a diverse multimedia content source, opening new opportunities and applications to explore and consume video content. A modular framework and algorithm for the representation engine and template collection is described. The representation engine based approach is evaluated through development of a prototype system grounded on the design of the proposed approach, allowing users to perform multiple content exploration tasks within a video. The evaluation demonstrated RAAVE's ability to provide users with a more engaging, efficient and effective experience than a typical multimedia player while performing video exploration tasks.

Keywords: multimedia analysis; video representation; multimodal video processing; human media interaction

1. Introduction

This study is based on a simple premise, namely, that despite all the research on video content, the general view of video content is that it is a sequence of moving images with or without an audio component. It is our contention that such a constrained view limits the potential of video as a content source. Therefore this paper explores the idea of viewing video content as a diverse multimodal content source, opening new opportunities and applications for exploring and consuming such content.

Content consumption is becoming increasingly video oriented [1–3]. Whether a person wants it for entertainment or for learning something new, one typically ends up relying on more video content than ever. Take YouTube as an example: over a billion hours of video content are watched daily [4]. In a white paper on global internet trends, CISCO estimates that video traffic will account for 82% of all internet traffic by 2021, up from 73% in 2016 [5]. The reasons for this are obvious. High-speed internet has made access to high-quality videos very convenient and new devices have lowered the barriers to publishing video content [6–8]. However, this ease of availability is not the only reason for the increasing reliance on video content.

Video is also one of the most versatile forms of content in terms of multimodality [9]. By multimodality, we mean that videos communicate content through multiple media, namely: the moving image track, the audio track and other derived features, such as a transcription of spoken words, translated subtitles, etc. Together, these modalities can provide an effective way of communicating information. The content value of these modalities as a whole, far exceeds their separate values.

Richness, both in terms of modalities and the amount of available video content, creates a challenge. Due to the high volume of video content available, it is becoming increasingly difficult for users to get to the relevant content with respect to the context or immediate search need. A recent study by Ericsson reports that an average American spends more than a year, over their lifetime, looking for something to watch on TV [10].

However, finding the right video among many is just part of the problem. As videos vary in length (up to several hours long), it is possible that a viewer does not need to consume the whole video, particularly if it is several hours long. It is possible that only certain parts of a video are of importance or interest to the user. So it is not only important to find the relevant video, but also to verify if the whole video is actually important to the viewer, or only a portion [3]. To put it in a different way, it is desirable to find not only relevant videos, but also the relevant portions of a relevant video. Waitelonis and Sack [11] observed that relevance is a highly subjective sentiment of the user which is dependent on context and pragmatics.

In addition, there may be other characteristics that are important for effective video exploration. Users might wish, for instance, to explore video content on multiple devices with different form factors and modalities (e.g., a mobile device or a home assistant device without a visual interface).

In essence, the increasing quantity and variety of video content, its mass availability and the proliferation of differing 'always-on, always-connected' devices, are creating new scenarios in which a user might consume video content. These new scenarios bring new challenges and opportunities with them. As will be elaborated in Section 2 that current video exploration approaches, while providing interesting use cases, are limited in fully harnessing the exploration potential of video content. To provide users with an effective exploration experience, a better approach might be to utilize the multimodality of video content in its representation and provide users with:

- The relevant content (the relevant portion of the video)
- The right manner or modality (due to device or personal preference)
- The right amount of detail (due to time constraints or personal preference)
- The segments surrounding the segment of interest (to get a better idea of the narrative)

1.1. Video Exploration

Video exploration is a complex task which is defined as a combination of tasks such as video retrieval (searching for videos in a collection), video navigation (search within a single video) and video summarization (synopsis of a video) [8,12].

The presented paper is focused on the navigation and synopsis parts of video exploration.

An effective exploration experience may be characterized as an engagement that enables the user to efficiently explore a video to discover content relevant to their needs. However, O'Brien and Toms [13] observed that due to the complex nature of exploratory search, traditional measures of information retrieval such as efficiency and effectiveness are not adequate to evaluate an exploratory search approach. They consider engagement to be a key quality of the process. Interactive video exploration research (Section 2) stipulates the importance of flexibility in an exploration approach. Therefore, it is our contention that an approach to explore video content should not just be efficient and effective; it should also be engaging and flexibly interactive. In order to identify the relationship with user engagement, it is important to first define user engagement with video content in the current context. According to O'Brien and Toms [13] user engagement is based on six factors such as

Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN) aspects of the experience. Specifically for videos [14,15] analyse user engagement by measuring for how long a user watched a video.

Questionnaires are also a very common method for analyzing engagement factors in video artifacts as in [16,17]. For this paper, the questionnaire used to assess user engagement with video exploration is based on the questionnaire proposed by Luagwitz et al. [18]. It is designed to compare the user experience with two systems and asks the users questions regarding the above-mentioned user engagement factors.

This paper describes an approach named An Alternative Representation of Video via feature Extraction (RAAVE) to represent the content of a video to users in order to enhance its exploration potential by providing a more engaging experience. The approach works in two phases. Firstly, state of the art tools are used to extract features along with timestamps from different modalities of the video stream. Then, upon receiving a content request, a representation engine utilizes a template collection to represent the content of a video in an appropriate configuration. A configuration determines the presence and granularity of certain features in order to compose a representation of the source video. A video can have multiple multimodal representations. Therefore, a representation may only have a subset of all available multimodal features.

In terms of novel aspects, this paper proposes a novel approach to automatically transform video from a linear stream of content into an adaptive interactive multimedia document, and presents a prototype system based on this approach to support automatic curation of video content on demand.

As stated above, this paper focuses on the navigation and synopsis aspect of the video exploration and not on the retrieval. For this reason the evaluation of the proposed video transformation approach focuses on the user ability to explore within a single video contrast to video searching tools which are designed to search over a large collection of videos such as the ones in Video Browser Showdown or VBS [12].

The proposed approach, RAAVE, is designed to be used in conjunction with a video search and retrieval system. The problem RAAVE addresses is, therefore, that of supporting the user in exploring a video to find specific content once a video retrieval system has retrieved video(s) from a large collection. As elaborated in Section 2, current SOTA approaches only provide the ability to either jump to a particular section of a video (non-liner navigation) or there are approaches that provide a smaller video extracted from the source video (video summarization) to let the user quickly skim the video. However, they are inadequate as in a typical exploration task user needs keeps evolving [1,19].

RAAVE takes a different approach and transforms a video into a multimedia document automatically so that the user can explore the content in different configurations and perform exploration related tasks (both navigation and synopsis) using RAAVE as per their evolving needs.

The following are the main contributions of this paper:

- an algorithm for our approach to the transformation of video content;
- a novel system based on the proposed architecture;
- a comparison-based user study to evaluate the effectiveness of the prototype system and the proposed approach in exploration tasks, including:
  - finding a particular piece of information within video content, and
  - evaluating the context of a video in a short time and writing a synopsis.
- In addition, this study assessed user engagement, comparing RAAVE to a baseline approach.

The rest of this paper is organized as follows. Section 2 described the state of the art. Section 3 described the design principal of RAAVE, while the details of the design of the proposed approach are described in Section 4. In Section 5 a prototype system based on the design is described. Section 6 describes the user study and the results of the user study which was conducted to evaluate the

proposed approach in multiple exploration tasks are described in Section 8. Section 9 provides a discussion of the results of the user study and finally Section 10 concludes the paper.

2. State of The Art

The exponential growth of video content has created challenges to explore video content effectively [1,3,20]. It is simply too distracting for the user to show a large list of videos in response to a query [1]. Researchers have proposed different techniques to mitigate the problem for example Waitelonis et al. proposed a semantic search approach, based on linked data to show relevant videos Waitelonis and Sack [11]. However, due to its multimodal nature, it is far more challenging to explore videos compared to textual documents [20] and in a recent survey paper Cobarzan et al. identified the importance of user interactivity in the video exploration process [12]. Therefore, researchers utilize multimodality and interactivity extensively in video exploration approaches.

2.1. Nonlinear Video Exploration

Many systems have been proposed which help users explore time-based media [21,22], and video more specifically, in a modular and non-linear manner. Barthel et al. [23] propose a collaborative approach that enables different users to create a video that provides alternative paths to navigate the content to learn about a topic. Merkt and Schwan [24] provide table of content style links to navigate to different sections of a video. This idea is extended by Pavel et al., in their study authors use a chapter/section structure to provide a skimmable summary and thumbnail of a video segment to a user [25]. Similarly, Meixner and Gold [26] design and evaluate an approach to create a table of content structure to non-linearly navigate a video for smart-phones and tablet devices. A widely used approach to enable users to non-linearly explore video content is creating hypervideos.

2.1.1. Hypervideos

Hypervideos are based on the same notion as hypertext i.e., hypertext ideas applied on a video [27] with earliest method for video branching proposed as early as 1965 [28]. Boissiere propose a system to identify topic changes in news broadcast by searching for special characters like '>>' placed by transcribers in news transcript to segment the video and provide hyperlinks to the identified segment [29]. This idea is enhanced by Finke and Balfanz in which authors propose a modular architecture for a hypervideo system for interactive TV portal that consists of an annotation engine, hotspot identification, a metadata format and presentation engine [30]. Stahl et al. apply the idea of hotspots and link nodes in educational settings and extend it with the ability to link additional material such as external web pages etc. [31]. The idea of supplementary material is used by Hoffmann and Herczeg in their study, the authors use hypervideo principle to create a personalized and interactive storytelling system which consists of a customized video player [32]. Aubert et al. extends the idea of storytelling by hypervideo with the use of structured metadata schemas such as RDF annotations [33], similarly RDF annotation combined with automatic entity extraction is proposed by Hildebrand and Hardman [34] to generate annotations for interactive TV programs.

Interactivity is an important aspect of hypervideos. Leggett and Bilda [35] experimented with alternative designs to allow user to navigate a hypervideo by reference images or a line following or grid etc.

Shipman et al. devised an approach to automatically create navigation links for hypervideo by proposing a hierarchical summary generation method to provide detail on-demand video content browsing [36]. In order to generate hierarchical summaries authors used low-level multimodal features such as color histograms and closed captions, clustering algorithms and heuristics depending on the genre of videos to segment video clips and determine the number of levels for the hierarchy and generate hyperlinks for navigation. Authors also designed a custom interface to search the collection and a specialized video player to browse the content. Tiellet et al. use the idea of detailed on-demand in an educational setting by offering a multimedia presentation of content [37]. In their study, authors

use the hypervideo system which offers links to more detailed information to students in the form of high definition images and supplementary text and annotation to learn surgical procedures.

Sadallah et al. [38] observed that prior hypervideo approaches were based on ad-hoc specifications and hypermedia standard such as SMIL [39] and NCL [40] are not well-suited for hypervideos. Authors propose an annotation-driven and component-based model for hypervideo inspired by other multimedia standards but more suited for hypervideos. While Mujacic et al. [2] proposed a different approach, in their study authors propose to use an hypervideo generation approach based on SMIL [39] specification. In order to simplify the process of authoring hypervideos, Meixner et al. propose an authoring system to allow non-technical users to create XML based annotations for hypervideo systems Meixner et al. [41]. Girgensohn et al. [42] use a hypervideo system and collaborative annotation to offer dynamically generated links to consume the content of meeting recording asynchronously.

While the above-mentioned approaches create hypervideo using video content, Leiva and Vivó [43] took a different approach. In their study authors use web page interaction logs with a web page to synthesize an interactive hypervideo to allow a user to visualize webpage usage. Similarly, Petan et al. [9] propose a similar approach to synthesize interactive hypervideos for corporate training scenarios.

In recent survey papers, Sauli et al. [27], Meixner [44] define the following as the primary aspects of all hypervideos base approaches.

- An authoring environment for annotations and setting up navigation paths,
- A meta-data structure for annotated data and navigation links,
- A specialized environment including but not limited to a customized video player for consuming the hypervideo.

Both Meixner [44] and Sauli et al. [27] consider the complexity of hypervideo systems both in terms of production (i.e., authoring systems) and consumption environments, to be an issue that is affecting the value of such systems in exploration tasks.

While hypervideos give more flexibility to the viewer in consuming the content, the flexibility is still limited to the extent to which the author has anticipated it. The viewer cannot go beyond that and while hypervideos do provide means to consume information in a multimodal manner. The multimodality comes from additional artefacts embedded by the curators instead of utilizing the potential of video as multimodal content source.

2.2. Video Summarization

To allow users to get the essence of a video in a shorter time, researchers have proposed many approaches which are referred to in literature as video summarization. Video summarization is defined as a technique that facilitates video content consumption by extracting the essential information of a video to produce a compact version [45]. It would not be farfetched to say that in video summarization, importance is usually attributed to visual features. For example, Benini et al. propose an approach to build a video summary or video skim by Logical Story Units (LSU) [16]. They create LSUs with salient features such as motion intensity in MPEG I-frames and P-frames and face detection in frames trained over Hidden Markov Models (HMM). De Avila et al. create a video summary by extracting color features from frames trained by unsupervised clustering by k-means clustering algorithm [46], while Almeida et al. use color histograms in I-frames of MPEG encoding and a noise filtering algorithm to a generate summary of videos [47]. Zhang et al. try to create a multi-video summary of user-generated videos based on aesthetic-guided criterion [48]. Belo et al. extend the idea of clustering key-frames by proposing a graph-based hierarchal approach [49].

However multimodal features are also getting considerable attention due to the added value they bring. For example Chen et al. use both visual and audio features to propose a hybrid approach combining content truncation and adaptive fast-forwarding to offer users a summary of video [50]. Kim et al. use low-level multimodal features and a fusion algorithm to create clusters that are

then utilized to create video summaries [51]. Wang et al. utilized multimodal features to develop an approach to segment program boundaries by getting program-oriented informative images (POIM) [52]. They then used these POIMs as basis to get keyframes and representative text to offer visual and textual summaries of the segmented programs. Hosseini and Eftekhari-Moghadam use multimodal features and fuzzy logic-based rule set to extract highlights of soccer games [53]. A comprehensive multimodal feature extraction can be seen in Evangelopoulos et al. [54]. In which authors take advantage of all three visual, audio and linguistic modalities and different data fusion techniques to create video summaries. The idea of multimodal fusion is further investigated by [55] while [56] utilize deep learning methodologies with different levels of explicit super-vision for learning attention maps for generating image description.

The aim of video summarization is essentially to create a new video artifact from the source which is shorter in length. In video summarization, the applied methodologies always choose the content to be included in the output artifact autonomously. The user is not part of the decision-making process and is only shown the final output. This makes it different from the proposed approach because it aims to present the multimodal content extracted from the source to the user and empowers him/her to choose the content they need either to consume immediately or to create a custom multimodal artifact of their choice for later consumption.

2.3. Gaps in the State of The Art

The state of the art review has revealed that while current approaches do provide interesting applications, they are limited in utilizing the potential of video while representing the content. For example hypervideos and interactive video navigation systems do allow users to explore video content in a nonlinear and interactive manner and there have been some attempts to allow users to explore content at different levels of detail [36,37,57,58]. Multimodality of video content is still underutilized in the presentation of content to the user. In essence, the state of the art approaches are limited in either one or more of the following aspects.

- Lack of user control in the configuration of the representation of content
- Solution is either designed to provide an overall synopsis of the video or search for some particular and not a combination of both, which affects the user experience in tasks which have evolving exploration goals
- User's ability to interact with the content is either limited or the interface is designed to be either:
  - content-dependent
  - overly complex
- Require prior curation by humans i.e., manual effort.

It is our contention that an exploration approach can utilize multimodal features more widely to enhance the user's exploration experience with video content. The approach should automatically curate content on-demand by representing the content in a configurable manner. By configuration, we mean that content representation may be configurable not just in terms of the amount of detail but also in the choice of combination of different modalities. Automatic curation of extracted features would minimize the dependence on prior human curation and supplementary material, and would allow users to potentially get more value out of video content. While providing the ability to change the configuration of the representation, enable users to go beyond the anticipation of designers and customize the content to their evolving exploration needs.

3. Proposed Approach

This section describes the basic idea of our engine based approach. As described in Section 1.1 , video exploration can be defined as combination of different tasks namely:

- Retrieval
- Navigation
- Synopsis or summarization

The proposed approach RAAVE is designed to be used in conjunction with a video search and retrieval system and therefore focuses only on the navigation and synopsis aspects. As described in Section 2.3 current approaches underutilize the multimodality in content representation and granting user control in the level of detail of the information, in an interactive manner.

This is because current researchers approach the problem by creating highly customized interfaces that augment supplement informational and/or control elements around a video. The problem with such solutions is that they cannot change with evolving user needs thereby limiting the exploration experience of the user.

The proposed approach (RAAVE) solves this issue by taking a different strategy. Instead of supplementing information or customizing the end interface, the proposed approach represents automatically extracted multimodal features in a configurable manner, independent of the end user interface.

By configuration, we mean that content representation is configurable not just in terms of the amount of detail, but also in the choice of combination of different modalities. The configuration of extracted features in the representation is done with the help of templates. Hence, the representation of content can be modified by changing the template selection. In order to select templates for representing the content, the proposed approach utilizes a representation engine that uses a template collection and a template matching process. The template collection and template matching process are inspired by the findings of the experiment performed by Salim et al. [59]. The following section details the design of the proposed approach.

4. Raave Design

RAAVE utilizes the fact that a video is not just a single/homogeneous artifact but, it is a combination of different temporally bounded parallel modalities (visual, audio, and textual). As described in Section 2, current approaches to representing video content are customized for a particular use case such as in the work of Monserrat et al. [60] and cannot be reconfigured for evolving user needs. To solve this problem, the proposed approach works as a representation engine independent of a user interface. Figure 1 shows an overview of the approach. A video is segmented into smaller segments and different tools are used to extracted multimodal modal features from its different modalities (visual, audio, linguistic). The information regarding the feature segments is stored in a repository. Upon receiving a request from a UI system, The RAAVE engine utilizes a template collection and relevance function to generate an appropriate representation of the video. The reminder of this section describes, in detail, the design visualized in Figure 1.

The proposed approach works in two phases.

- Extraction and Indexing
- Representation through template matching

Both phases work independently. Extracted features are stored in a repository. Representation is done independently of the extraction so that the configuration of the representation can be dynamic and flexible, thus using different templates to achieve different representations and interactions.

4.1. Extraction and Indexing

Steps involved in the phase are following.

- Video Segmentation
- Multi-modal Feature Extraction

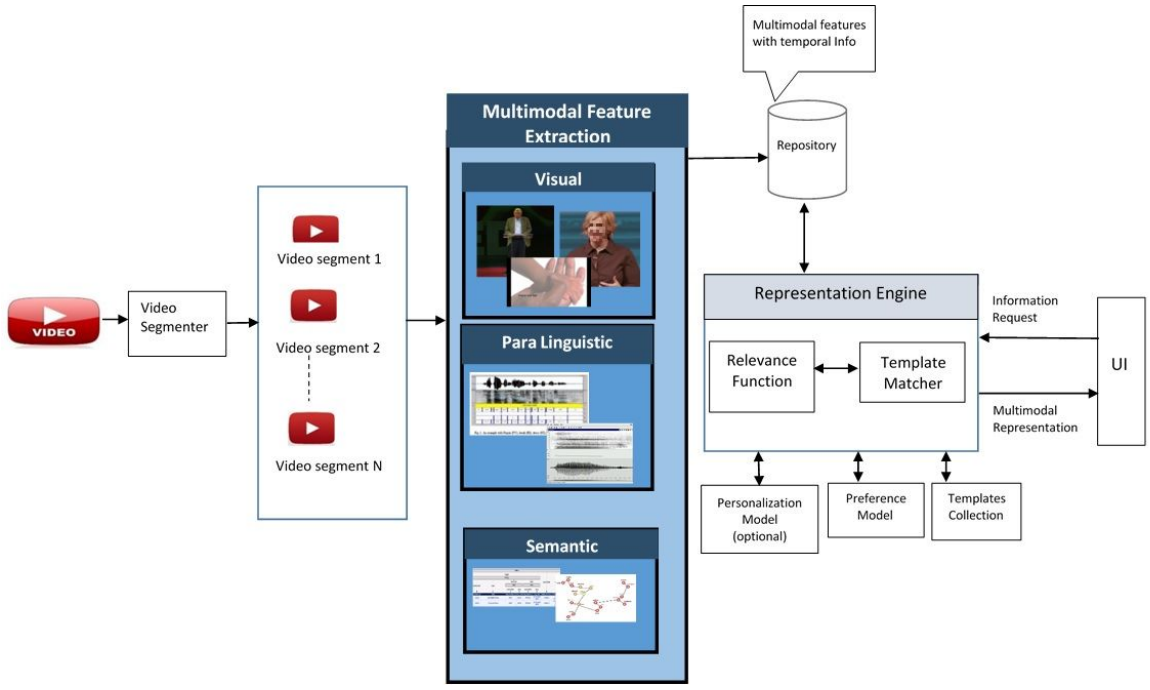

Figure 1. Overview of the proposed approach.

4.1.1. Video Segmentation

To generate an appropriate representation the proposed engine needs segments regardless of how they are segmented. The focus of this approach is not to devise a new segmentation technique but to utilize already existing segmentation techniques and then represent the segment in a multimodal and configurable manner.

Researchers have developed many techniques to segment videos (detailed in Section 2). The choice of a particular segmentation approach depends on many factors e.g., the genre of the video.

This paper is focusing on presentation style information video e.g., TED talks. Even with presentation style videos, there can be multiple ways in which a video can be segmented.

System can choose from already developed off the shelf techniques to segment a video into smaller units based on the genre of the video. As an example consider TED-style informational videos, the segmentation algorithm used in the experiment is C99 text algorithm [61].

Others segmenters which can be used are:

- visual
- multimodal and
- semantic etc.

4.1.2. Multimodal Feature Extraction

After segmenting the video, the next step is to extract multimodal feature from video segments along with their timestamp. In order to do that, the video needs to be decomposed into different modalities i.e., visual, audio, textual and video itself. State of the art tools and techniques can be utilized to extract features from these modalities.

By multimodal feature extraction, we mean the characteristics or features within the different modalities by which a video delivers its message to the viewer which may include the following.

- The visual modality i.e., anything visually interesting or engaging to the viewer, e.g., visual features like camera close up or visual aid, etc.
- The paralinguistic modality i.e., the audio features, e.g., laughter, applause and other audio features.

- The linguistic modality, i.e., the spoken words, also any text within the video as well as any supporting resources e.g., human written synopsis, etc.

Feature Expanse

The need for different features representation in certain configurations depends on the fact that different features have different expanse. Some features would offer more detailed content to the user i.e., they would have deeper expanse in terms of information value, while others would offer less detailed content to the user. However, the less detailed features would be more efficient to consume in terms of time.

For example, consider the actual video footage of a particular segment. It would offer the full content of that segment but it will require longer to see it compared to an automatic text summary generated from the segment transcript. The text summary would require less time for the user to consume but its expanse in terms of content value would be limited compared to the video footage. Similarly, consider key frames from the video footage or a word cloud of key terms from the textual transcript. Both will be efficient in terms of time but limited in terms of depth of information.

Hence different features may have different expanse of the depth of information and they would also belong to different modalities.

As mentioned above, the goal is to represent the content of the video in different configurations. By configurations, we mean the different combination of extracted features which intern would offer different expanse of information to users and they would do so in different modalities.

Feature Availability

It is possible that certain features are not present in a particular video. As an example, consider a TED presentation video in which the presenter does not use any PowerPoint slides or any other visual aid. For such a video, the tools designed to extract slides from a video would not return any output hence the keyframes feature for that video would not be available for that video which will affect the choice of potential representations for that video. For the sake of simplicity, the discussion onward considers two modalities only, i.e., visual and textual.

4.2. Representation Through Template Matching

Once the video is segmented and multimodal features extracted, the next step is to represent the segment in an appropriate configuration. To do that we are proposing a representation engine that utilizes a template collection (see Section 4.2.1).

For each of the segments of the video, the representation engine chooses a suitable template. A template is essentially a configuration setting to represent the extracted features.

The engine works on a request-response cycle. Upon receiving a request for information from a UI (User Interface), the engine does the following activities.

For each segment of the video it:

1. Determine the degree of relevance of the segment with the current request for information.
2. Based on the relevance, choose an appropriate template for representing the segment.

In order to perform the two tasks, the engine requires the following:

1. Determining the Relevance using a Relevance Function.
2. Template Matching by utilizing a Template Collection.

4.2.1. Template Collection

The Template Collection contains the list of templates for the engine to choose an appropriate template from.

A template basically determines which extracted feature or combination of features shall be included in the representation of a video segment.

A template has the following dimensions:

- Feature Expanse
- Primary Modality (see Section 4.2.2.3)

The number of templates are dependent upon the extracted features. As a template is a possible permutation of available features. Not all the permutations are included in the collection.

Table 1 lists the possible dimensions for template matching while Table 2 shows an example of a template collection.

Table 1. Dimensions of template matching.

	Dimensions	
	Expanse	**Primary Modality**
Possible Values	Efficient	Visual
	Deep	Textual Mixture

Table 2. An example of a template collection.

Template ID	Expanse	Primary Modality	Feature(s)
1	Efficient	Textual	Word Cloud
2	Efficient	Visual	Key Frames
3	Deep	Textual	Text summary
4	Deep	Visual	Video Snippet
5	Deep	Textual	Word Cloud, Text Summary
6	Deep	Visual	Key frames, Video snippet
7	Deep	Mixture	Key frames, text summary
8	Deep	Mixture	Word Cloud, Video snippet
		....	
9	Deep	Mixture	A permutation of extracted features.

4.2.2. Template Matching Criteria

The template matching process is based on the following 3 criteria:

- Relevance
- Expanse
- Primary Modality

Relevance

In order to determine what segment is relevant or important in a given context. The assumption is that relevant segments need to be represented in greater detail than non-relevant segments (see Section 4.4).

Expanse

Different features offer a different amount of information within a video segment i.e., they have different expanse. They are either deep or efficient. If a feature is efficient to consume, then it would not offer detailed content. Alternatively, if it is deep in terms of content then it would require more time to be consumed.

As an example of textual features take a word cloud of keywords generated from a transcript of a video segment. The word cloud is efficient to consume as it can glance quickly but it does not offer

much detail of what is discussed in the segment. Alternatively, if a text summary is generated from the transcript then it would take longer to read it however it will also give a more detailed understanding of the video segment than a word cloud.

Therefore, the word cloud is an efficient feature while textual summary is a deep feature of linguistic modality.

Primary Modality

A segment can be represented in either a single modality or a mixture of modalities depending on the context.

4.3. Template Matching

Template Matching process is essentially a 3-dimensional problem. Given a video to represent the engine had to find an appropriate template from the collection for each segment. The engine does so based on 3 criteria (detailed in Section 4.2.2).

Our assumption is that relevant segments would require a deeper exploration. Therefore, the first step in matching a template is finding the relevance value of a segment. Determining the relevance establishes the depth of the segment representation i.e., the choice of template.

4.4. Determining The Relevance

The proposed approach transforms video content based on the context. In order to choose a template for relevant segments, the representation engine needs to determine the relevance of each segment.

Whether a segment is relevant or not in a given context may be determined by many factors. For representation purposes it makes sense to assume that given a user query, if a segment contains the keywords of the user query, then it is relevant.

Apart from the query, personalized interest can also determine the relevance of a segment. A segment may be relevant if the user query terms appear or the user's topic of interest appears in the segment. Similarly, the segments surrounding the segment in question may also determine its relevance.

The following are some of the factors which may determine the relevance of a segment.

- Request context
- Personalization model
- Segments preceding and following the segment

4.4.1. Request Context

By request context, we mean the current information need of the requesting entity. It could include but not limited to the search query.

The context can also be information such as time, location or the device initiating the request.

4.4.2. Personalization Model (Optional)

There can be topics that the user might be interested in. So even if a video segment may not contain the info required by the current query request it might have info that might be of interest.

In case of the absence of a search query personalization model becomes more important to determine the degree of relevance of a segment.

4.5. Relevance Function

Relevance function is the component of the engine which takes as input a segment and relating factors and returns the relevance value of that segment.

$$Segment_Relevance = getRelevance(seg, Pseg, Fseg, pm, co).$$

where:

- Seg is the segment in question
- Pseg is the segment preceding the segment
- Fseg is the segment following the segment
- pm is the personalization model
- co is the request context

4.6. Choosing A Template

After determining the relevance of a segment what remains is, selecting an appropriate template for the segment. The representation engine must choose a template based on expanse and primary modality.

The engine determines the expanse of the template based on the segment's relevance i.e., deeper exploration is desirable if the segment in relevant.

Once the depth value has been determined, the only thing left to determine is the modality of the template. Now the modality can be singular (visual or textual etc.), or it can be a mixture. Based on previous experiments [59] we assume that mixed modality is appropriate for relevant segments. It is for this reason mix modality is used only for relevant segments and single modality for others.

The next step in narrowing down the choice of template is choosing a particular modality value to choose for the segment template. The engine must choose a modality value in the case of a single modality template. In case of mix modality, the engine has even more things to consider i.e., it needs to decide if all modalities will be represented by deep features or a combination e.g., visually deep and textually efficient etc.

The engine narrows down the choice of modality based on two criteria. They are:

- Segment Suitability
- Preference Model

Segment Suitability

As discussed in Section 4.1.2 multimodal features are extracted from video segments using different tools. It is possible that certain features were either not extracted from the segment for any reason or they are not suitable from the point of view of content value.

In a previous user study [59], it was found that users may prefer a particular modality i.e., textual information because he/she is a fast reader or they may prefer visual information. The representation engine may utilize the preference model to narrow down the choice of template for the segment.

In summary, the engine determines the modality by the combination of the segment's feature suitability and use preference model. It gives priority to segment suitability and if suitable, it uses user preference to narrow down the choice of template to represent the segment.

The pseudo-code for the representation engine is presented in the following section. This pseudo-code is used to develop the prototype used in evaluating the proposed approach.

4.7. Pseudo Code

Preconditions

1. Video has been segmented
2. Features are extracted and indexed
3. Template collection, preference/personalization model is available

For each segment do:
Determine if segment is relevant to a given context by relevance function.
If segment is relevant choose deep templates
Else choose efficient template
If segment is efficient then choose single modality
Find segment suitability for modality
If only one suitable template to choose from than represent segment
Else see user preference
Choose template with users prefer modality
Else choose mix modality
Choose visually efficient and textually deep or vice versa based on segment suitability and
User preference giving priority to segment suitability.

4.8. Design Summary

Effective exploration of video means that the user may not want to or need to consume all the video. To put it in another way, not all parts of the video may be equally relevant in each context. Therefore, it would make sense to show in the representation, the relevant parts of the video using more detailed representation. However, the user may not want to completely ignore the other parts of the video, to better comprehend the video, or for any other reason. Therefore, it would make sense to represent those parts in a less detailed representation. That way the user can focus on the relevant parts of the video without completely losing the information in the other parts.

To represent the segments to the user, RAAVE utilized a representation engine. The representation engine is essentially the main component of the whole approach. It is the link between the User Interface (UI) and the extracted features of the video segments.

In summary, the engine determines the modality by a combination of segment's feature suitability and use preference model. It gives priority to segment suitability and if suitable it uses user preference to narrow down the choice of template to represent the segment.

In short:

- a segment of the video is either relevant or not in a given context
- a segment is represented by feature(s) which are deep or efficient
- the representation belongs to modality visual, textual or mixture.

Hence, we can describe effective video representation as representing each segment based on the appropriate value of three dimensions which are expanse, relevance, and primary modality.

5. Prototype System

In order to evaluate the proposed approach, a prototype based on the design described in Section 4 is needed to represent the content of the video to users. Following is a description of the system developed to evaluate the proposed approach.

As per the design mentioned in Section 4 the prototype system works in two main phases.

5.1. Extraction and Indexing

5.1.1. Segmentation

A video can be segmented in many ways utilizing different modalities (detailed in Section 2). The choice of modalities is often domain dependent.

For the current study, the video is segmented by utilizing the textual modality. TED videos come with text transcripts in the SubRip subtitle file format, .srt, which is a widely supported format. In this format the subtitles are stored sequentially along with the timing information.

Thus, the transcripts of TED videos were first split into sentences using StanfordNLP toolkit [62] and fed into the C99 text segmentation algorithm [61] in order to produce video segments.

After the segmentation performed, the time information in the transcript file was used to determine the start and end time of each video segment.

5.1.2. Visual

To extract keyframes of a segment. A custom tool was developed using openCV [63] to detect camera shot changes. From those scene changes, one frame from each shot was selected. From those selected frames, frames with and without a face were identified using a HAAR cascade [64]. Speakers often use visual aids, such as presentation slides in a TED presentation. The heuristic was that shot without a face after a shot with a face might contain some images of visual aid used by the presenter which could contain useful information. Users can tap or click on the frame to see all the selected frames to get a visual synopsis of a particular segment.

5.1.3. Summary

Automatic text summary is generated from the transcript of the segment. An online summarization tool for generating text summaries [65] was utilized.

5.1.4. Key Terms

Word cloud generated from the transcript of the segment. We used the online tool TagCrowd [66] for the word cloud generation.

5.1.5. Video

The text in the transcript comes with timestamps. Once the segmenter segmented the text, timestamps were used to determine the start and end time of a particular segment and its video snippet was offered to users to watch.

5.1.6. Indexing

For the representation engine to offer users the extracted features, the features along with their timestamp information need to store in an efficiently retrievable manner. To do this all the data related to video segments and the multimodal features along with the timestamp are stored as documents in a tables or cores in Solr search platform [67]. Solr is written in JAVA, uses Lucene indexing and searching engine [68]. A simple web request is used to access the information which is served in standard machine-readable format such as JSON (JavaScript object notation). Following is an example of a video segment indexed by Solr.


```
{
"id": "ThomasPiketty_2014S-480p_c99_2",
"video":["ThomasPiketty_2014S-480p"],
"num":[2],
"segmenter":["C99"],
"Start_time":["00:02:42"],
"End_time":["00:07:58"],
"seg_text":["\n\tSo there is more going
on here,
but I'm not going to
talk too much about this today,
because I want to focus on wealth
inequality.
So let me just show you a very simple
indicator about the
income inequality part.
So this is the share of total income
going to the top 10 percent
```

```
... "],
```

Similarly, information about multimodal feature is indexed for the representation engine to quickly search it. Following is an example of a multimodal feature indexed in Solr.

```
{
"id":"ChrystiaFreeland_2..." ,
"video":["ChrystiaFreeland_2013G-480p"],
"segmenter":["C99"],
"Expanse":["Efficient"],
"Modality":["Visual"],
"Name":["Keyframes"],
"seg_num":[1],
"FeatureValue":["CF/CF_1_vis.html"],
"_version_":1579238985017851904}
```

5.2. Representation Through Template Matching

5.2.1. Representation Engine

The representation engine is implemented as a server application developed using ASP.net MVC framework. The server-side works on a request–response cycle. The engine receives a standard HTTP request for video content in the form of a query and it responds by sending the multimodal data to the requesting application.

The representation engine does that by implementing the pseudo-code of Section 4.7 in C# language.

5.2.2. Template Collection

Templates are configurations of available multimodal features. In theory, any permutation of available features can be represented (detailed in Section 4.2.1). However, not all permutation may make sense for a representation. Therefore, an implementation may not have some possible templates in the collection. Furthermore, an implementation may only include a few permutations because of application design choices.

Template collection can be implemented in any format depending on the technology used.

For the prototype, the template collection is embedded in the representation engine source code since the design of the user study (Section 6) only required a subset of possible permutations of feature set.

The feature set entailed within this implementation included the following (see Section 5.1 for details):

- Extracted Keyframes from the video recording of the segment.
- Text transcript and textual summary generated from the transcript.
- Word Cloud (generated from transcript).
- Video recording of the segment.

This enabled the implementation of template matching as a series of simple if-else statement. The following code snippet shows two of the if-else branches implemented in C# language.

```
public class Template
{
String templatID;
ExpanseName expanse;
ModalityName modality;
List<FeatureName> featureset;
public string TemplatID
{
get { return templatID; }
set { templatID = value;}
```

```
}.... }

Feature PrimFeature, SecondFeature,
defaultFeature, ohtefeature;

if (videoSeg.Relevant == true &&
userPreferce ==
ModalityName.Textual && hasSummary
== true)
{
PrimFeature = videoSeg.AvailableFeatures
.Where(F => F.Name ==
FeatureName.Text_Summary)
.FirstOrDefault();
PrimFeature.RepPlace = RepresentationPlace
.Primary;
segRep.FeatureSet.Add(PrimFeature);
.....

else if (videoSeg.Relevant == false && userPreference ==
ModalityName.Visual && hasKeyFrames == true)
{
PrimFeature = videoSeg.AvailableFeatures.
Where(F => F.Name == FeatureName.Keyframes).FirstOrDefault();
PrimFeature.RepPlace = RepresentationPlace.Primary;
segRep.FeatureSet.Add(PrimFeature);
```

The above code snippet shows the implemented template selection. As per the pseudo-code in Section 4.7, relevant segments are represented with deep features like text summary/transcript or video recording of that segment, depending upon the user preference for modality while non-relevant segments get efficient features such as word cloud or key-frames depending upon the user preference for modality.

The actual placement of the represented features and any other information is dependent upon the user interface. Section 5.2.4 describes the implemented user interface.

5.2.3. Determining Relevance

In the prototype, the relevance of a segment is determined solely based on the query request. Once the representation engine receives the request it passes the query to Solr system [67] which returns segments as relevant, in which the keywords of the request query appears.

5.2.4. Video Search System

The prototype is built on a query-based video search. The prototype is essentially a web application that allows users to search for videos based on their queries. The first web page is a simple query box where users can enter their query text and search for relevant videos.

Once the user enters a query the next page shows the list of videos relevant to the query. Just like any video service, it gives the title and a small description of each video. The user can click on the video to explore its content.

Figure 2 shows the automatically generated representation of the content of the video by the prototype engine. Users can use the provided search box to search for information within the video. The video is divided into segments and each segment is represented based on the value of relevance function (Section 4.5) which currently is a binary value. For relevant segments, the primary representation area shows features with deep expanse while for non-relevant segments efficient features are placed in the primary representation area. Depending on user preference which can be selected by the preference buttons: the primary representation area shows either visual or textual features.

The developed prototype is used to perform the experiments to evaluate the proposed approach. The following section describes the experiment in detail.

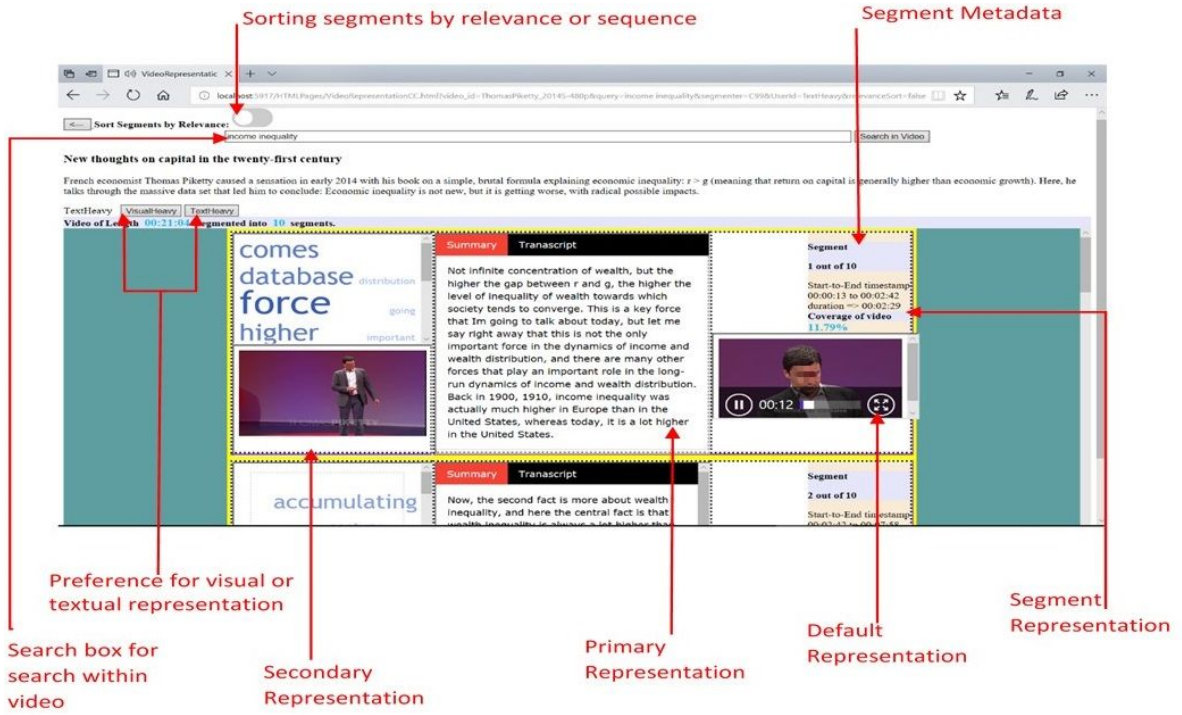

Figure 2. Video representation by representation engine.

6. User Study

The goal of this experiment is to evaluate the proposed approach (detailed in Section 4). The evaluation is done by conducting user studies, utilizing a prototype implementation of the proposed approach (detailed in Section 5). The experiment is performed as a comparison study to evaluate the performance of the proposed approach named RAAVE to a baseline system which is a standard video player.

Hypotheses

The main hypothesis of the experiment is "RAAVE engine provides an enhanced exploration experience to the user compared to a baseline video player by enabling efficient and effective video navigation, synopsis and better engagement".

In order to evaluate the main hypothesis i.e., exploration experience, it is divided into three sub-hypotheses as per the discussion in Sections 1 and 2.3, following lists the sub-hypotheses.

- **Hypothesis 1.** *Users have a better experience interacting with RAAVE compared to the baseline player.*
- **Hypothesis 2.** *RAAVE is better at allowing users to search for information in different parts of a video compared to baseline.*
- **Hypothesis 3.** *Users can quickly get a better understanding of the content of video using RAAVE compared to the baseline player.*

7. Experiment Design

In order to evaluate the above hypotheses, the experiment is designed as a comparison study between two systems (RAAVE and Baseline). Participants "users" performed two types of tasks; the answer search task for the evaluation of Hypothesis 2 and the synopsis writing task to evaluate Hypothesis 3. After performing the tasks users were asked to give feedback about their experience with both systems. This was done to evaluate Hypothesis 1.

This paper is about evaluating the extent to which multimodal features can enhance the exploration experience of users within video content. In order the enhance the experience this paper

has proposed a template-driven representation engine that represents multimodal extracted features in different configurations. The goal of this experiment is to test the proposed representation engine with users performing exploration tasks with video content. The proposed approach does not propose a user interface (UI).

The representation engine is designed to be UI agnostic as the end interface may be customized for the end-user device and other factors. However, a UI is needed to perform the experiment as users need it to interact with the represented feature set. The experiment is to evaluate the user experience with automatically created documents, not a particular user interface and compare it to the baseline video player which only plays video. Therefore, the user interface for the prototype is designed to be minimalist and bare bone so that it is the approach that gets tested and not a UI.

Apart from evaluating the hypotheses, another goal of the experiment was to assess user behavior while performing different tasks. By user behavior, we mean the usage patterns of users with the different modalities and feature set while performing searching for answers for some particular piece of information or trying to get the overall synopsis quickly. For this reason, the experiment is designed to stress test the RAAVE system.

7.1. Expected Outcomes

- Evaluation of the set of hypotheses (details in Section 6).
- Usage patterns with both systems.
- Differences in user interactions with the two systems while performing the tasks.
- Are there particular feature representations offered by RAAVE which could be more useful than others?

7.2. Systems

7.2.1. Raave System

RAAVE system is the system based on the proposed approach described in Section 4. The details of the system are described in Section 5.

7.2.2. Baseline System

To evaluate the hypothesis Section 6. RAAVE is compared with a baseline system which in this case is a simple video player. Admittedly, the choice of using a simple video player as a baseline is an unusual one. Traditionally researchers compare their approach with an approach from the state of the art. However, it was not feasible for the current experiment. The proposed approach (RAAVE) automatically transforms video into a multimedia document i.e., RAAVE transforms video into something more than a video. State of the Art (SOTA) approaches that do that are Hypervideo based approaches (detailed in Section 2.1.1). As detailed in the SOTA review (Section 2), Hypervideo systems require human curation by utilizing specialized authoring environments and video players. Since RAAVE utilizes automatically extracted multimodal features, therefore, it is not feasible to have a direct comparison with Hypervideo systems due to this fundamental difference in the approaches. Moreover, while Hypervideo systems do represent multimodal information with video content, they utilize supplementary content from sources other than the source video, while the goal of the RAAVE approach is to utilize the content within the source video in novel ways to enhance the user experience thereby making a direct comparison not feasible due to the fundamental difference in the approaches.

Another reason to choose a simple video player as a baseline goes to the main goal of the presented paper. As described in Sections 1 and 2 despite all the research; the current view on video content is essentially a continuous stream of images with or without an audio component. As RAAVE proposes to consider video content as a diverse content source by transforming it automatically, it is natural to compare the transformation based approach with the continuous stream of images with a parallel

audio component i.e., a regular video player as a baseline which is also ubiquitous due its familiarity and common usage.

Figure 3 shows the baseline system which users utilize to interact with the video. It contains a standard video player with pause/play button and a scrubber so that user can drag it across to watch a different portion of a video. Underneath the video player is information regarding the start and stop time of different segments of the video. It is provided to aid the user in performing the answer search task.

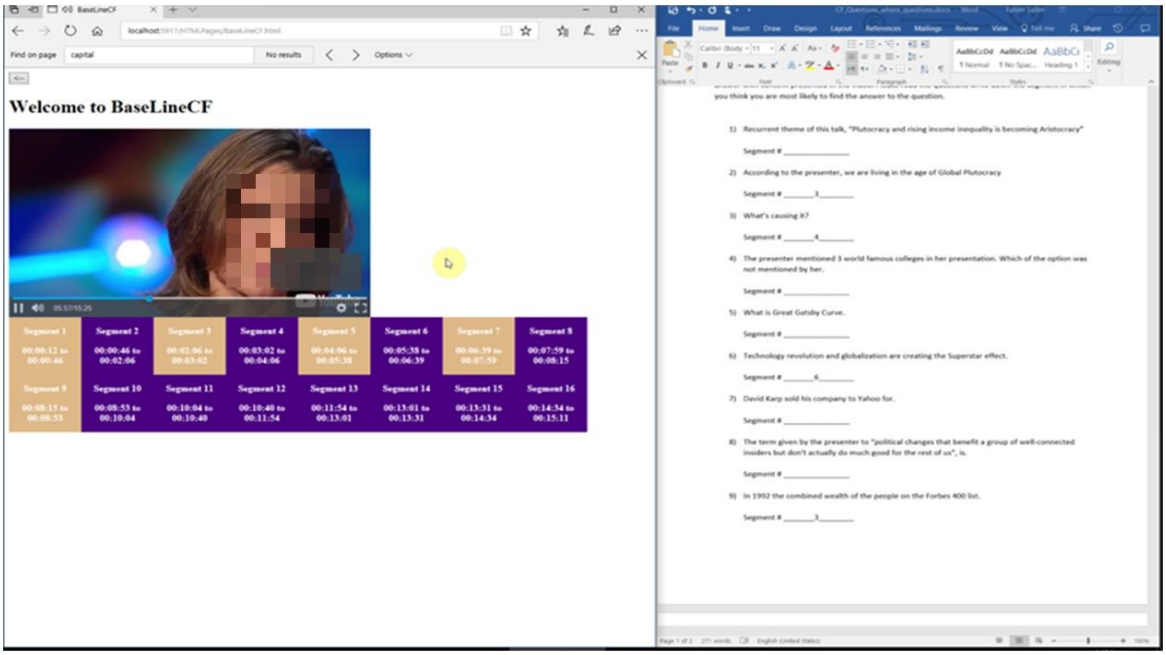

Figure 3. (**Left**) Baseline video player and information regarding the start and end times of segments of the video. (**Right**) Answer search task using baseline system (screen shot).

7.3. Tasks

The following are the two tasks performed by users in the experiments.

7.3.1. User Experience Questionnaire for Engagement Assessment

In order to evaluate Hypothesis 1, i.e., comparison of user experience with the two systems, after performing the experiment tasks, users were asked to fill the user experience questionnaire [18]. The User Experience Questionnaire (UEQ) is designed to compare the user experience with two systems. Users are asked to fill a questionnaire consisting of 26 questions. Each question consists of a pair of contrasting attributes and seven-point Likert scale between them. Figure 4 shows three of the 26 questions in UEQ.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3

Figure 4. Example questions User Experience Questionnaire (UEQ) first 3 of 26. (For full list see Appendix A.1).

The 26 questions of the UEQ can be categorized in to following categories.

- Attractiveness
- Perspicuity
- Efficiency
- Dependability
- Stimulation
- Novelty

The user filled the UEQ twice, once for RAAVE and once for the baseline.

7.3.2. Answer Search Task

The answer search task is designed to evaluate the ability to find information within a video. Users are given a set of questions that have to be answered by utilizing the content of the video as quickly as possible.

The answer search task was performed by 12 users. They performed it on all four videos for a total of 24 times. Each user performed the task twice, once using the proposed system and once the baseline system. The order of the system and the video were always changed i.e., some user performed the task using RAAVE first and baseline the second time while others did it vice versa. Following is an example of questions for video [69]. (For full list of questions see Appendix A.2).

```
1.      Jane Austen is mentioned in.
Segment # _____________

2.      What made the Swiss show flexibility in bank secrecy?
Segment # _____________
```

For each video, there are 14 questions. Figure 3 shows a screenshot of an answer search task using the baseline system.

Since hypothesis B is about comparing the ability to search for information within a video, therefore in this task, instead of providing the actual answer to the question users were asked to provide the segment number which contains the information needed to answer the question that identifies the portion of the video in which they think contains the relevant information.

For example, consider the question "What made the Swiss show flexibility in bank secrecy?" The user simply tells the number of the segment in which the information to answer the question appeared i.e., segment # 7. Both RAAVE (Figure 2) and Baseline system (Figure 3) contained information for the user to easily identify the segment number.

7.3.3. Synopsis Writing Task

The second task corresponds to the goal of enabling the user to get the essence of the video effectively. This task is designed as a comparison study to evaluate the user's ability to get the essence of video effectively which is to evaluate Hypothesis 3. Each of the 12 participants performed the synopsis writing task twice, once using the RAAVE system and once the baseline player. Figure 5 shows an example of a synopsis writing task using the RAAVE system.

In this task, participants consume the content of video for a shorter amount of time compared to the length of the video and write a synopsis of the video. Throughout the video summarization literature (Section 2.2) researcher has used the ratio 0.2 to test their video summary generation approach.

Therefore, this experiment also uses 0.2 as the ratio for the amount of time given to the user to consume the content of video so that they can write a synopsis. For example, for a video of 15 min user could consume the content for up to 3 min. Note that this is not the amount of time to write the synopsis but to consume the content of the video, users were allowed to take as much time they wanted to write the synopsis. It was left to the user's discretion if they wanted to start writing the synopsis or take notes while they were consuming the content and continue writing the synopsis after the allowed time passed or they consume the content first and write the synopsis later.

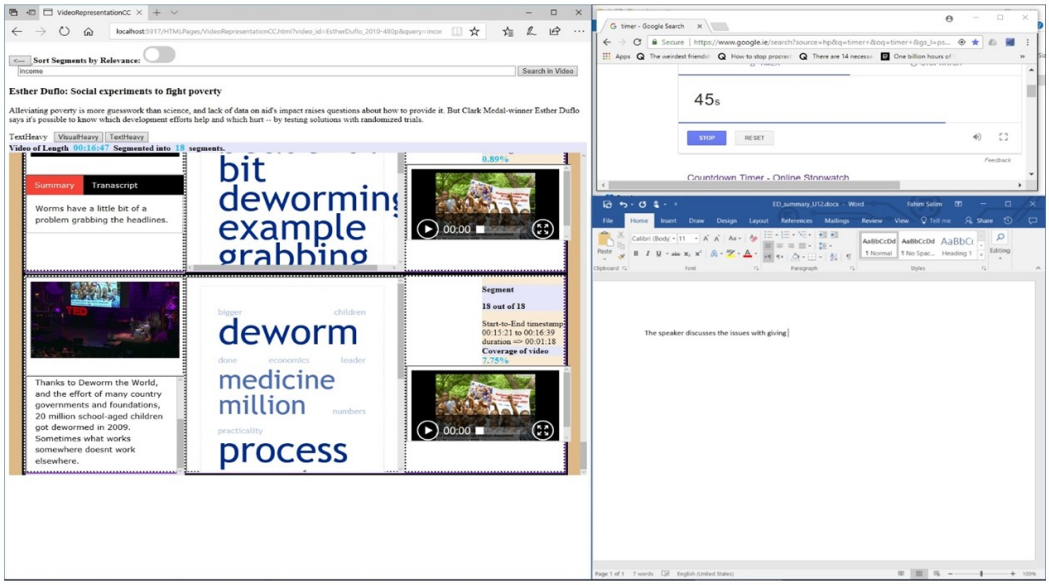

Figure 5. Synopsis writing task using An Alternative Representation of Video via feature Extraction (RAAVE) system (screen shot).

To compare user performance with the two systems, another set of participants "reviewers" were asked to evaluate the synopsis produced by the users (details in Section 7.6.2).

Hence the experiment has two systems, two types of tasks and two types of participants. Table 3 summarizes the configuration.

Table 3. Experiment items.

Systems	Participants	Tasks
RAAVE	Users	Answer Search
Baseline	Reviewers	Synopsis writing

To summarize: participants (users) use two systems (RAAVE and Baseline) so that their performance could be compared for the two tasks (Answer search task for Hypothesis 2 and Synopsis writing for Hypothesis 3). In order to compare their performance users were asked to perform both tasks twice. Each individual user session had 4 attempts. Table 4 lists the attempts for two users an as example to make things clear.

Table 4. Attempts by users.

User	Attempt 1	Attempt 2	Attempt 3	Attempt 4
1	RAAVE system to perform Answer search using (video 1)	Baseline system to perform synopsis writing (video 2)	Baseline system to perform Answer Search (video 3)	RAAVE to perform Synopsis writing. (video 4)
2	Baseline system to perform Synopsis writing (video 4)	RAAVE system to perform Answer search using (video 3)	Baseline system to perform Answer Search (video 2)	RAAVE to perform Synopsis writing. (video 1)
3	..	..	..	..
n	..	..	..	..

Since each user performed four attempts, the experiment uses 4 TED talks. The following section describes the test videos.

7.4. Test Videos

For the user study, a total of 4 TED videos were utilized. The number was chosen to ensure that each user explores a different video for each of the experiment attempts. TED videos are chosen due to their general-purpose nature and appeal to a wider audience therefore it made the selection of experiment participants simpler as any person familiar with informational style videos such as TED was qualified. All sample videos belong to the same topic, namely "Economic Inequality". This topic was chosen due to the fact that none of the study participants had an educational background in economics. It is to ensure consistency in the experiment tasks [1,7].

While TED presentation videos have a consistent structure [70] there can be slight variations. For example some presenters use visual aids while others prefer to talk without any slides. Similarly, some presentation ends with a supplementary item such as an interview etc., while others only consist of a presentation. The sample video was chosen to be representative of the general TED video collections. Two of the TED videos used contain visual aid i.e., the presenters use slides and pictures in their presentation [69,71] while the other two presenters do not use any visual aid [72,73]. The reason for choosing two videos with slides and the other two without slides follows the idea of content value of different modalities (Section 1.1). For example in the video ED [71], the presenter speaks about mumps causing deaths in NY. She never tells the listener about the total number but a slide in her presentation shows the number. Now a question related to this can only be answered by utilizing the visual modality. Listening to audio-only or running a text search on the transcript of the presentation would not yield the answer. In the researcher's opinion, videos without visual aid were relatively easier to comprehend than the other two.

In TED presentations, the presenter often presents the main idea of the presentation at the beginning of the video as in Freeland's presentation [73], the presenter gives the main idea of the presentation early on and gives some details to reiterate it. This makes it easier for users to get the overall synopsis because users can still get the main idea of the video even if they did not consume all the portions. Whereas Collier and Duflo [71,72] describe a problem and then offer a solution later and summarize the discussion in the end. Therefore, users are more likely to miss the essence of the presentation if they do not consume all the portions compared to the first video. The 4th video [69] is an interesting case. While the presenter does give the main idea in the beginning, the technical nature of it makes it difficult for viewers to fully get the point, especially if the viewer does not have an economics background, it is only by watching the middle and the end parts of it that a viewer can get the essence, making it a relatively difficult video to comprehend.

By choosing videos with and without visual aid and easy and difficult videos, the exploration experience of users with different types of content can be evaluated.

7.4.1. "New Thoughts on Capital in the Twenty-First Century" by Thomas Piketty

This TED talk [69] is approximately 21:00 minutes in length and consists of two parts: the first part consists of a presentation while the second part is an interview. The presenter used slides and charts extensively during the presentation. It is our opinion that the information presented in the video is on average more technical than a typical TED video. Throughout this paper, this video would be referred as TP.

7.4.2. "The Rise of the New Global Super-Rich" by Chrystia Freeland

This TED video [73] is approximately 15:20 minutes in length and it is different than the first video in a number of ways. Firstly, the video solely consists of the presentation and contains no interview, furthermore, the presenter does not use any slides or any other visual aid during her presentation. It is our opinion that this video was easier to comprehend than the first one due to the general nature of the information provided. Throughout this paper, this video would be referred as CF.

7.4.3. "Social Experiments to Fight Poverty" by Esther Duflo

This TED video [71] is approximately 16:40 minutes in length. The presenter uses slides and charts extensively during the presentation. It is our opinion that this video while not as technical in nature as [69] does contain a lot of information. Throughout this paper, this video is referred to as ED.

7.4.4. "The Bottom Billion" by Paul Collier

This TED video [72] is approximately 16:51 minutes in length. The presenter does not use any visual aid during the presentation. In a manner similar to [73] this presentation, in the researcher's opinion does not contain too much technical information and is easy to comprehend for a general audience. Throughout this paper, this video is referred to as PC.

7.5. Experiment Participants

The experiment has two types of participants.

- Users
- Reviewers

7.5.1. Users

A total of 12 (each performed the experiment 4 times) users performed the two tasks in the experiment, 7 males and 5 females. The design of the experiment allowed us the use the same person multiple times. Traditionally in comparison studies, the users are split in two half i.e., experiment group and control group. However, since each user in our experiment performed four tasks with different systems and different videos and in a different order each time (see Table 4 for details) resulted in 48 trials.

A sample size of 12*4 (effectively 48) users seems adequate compared to studies on this subject [41,74]. All the participants in our experiment have postgraduate degrees in computer science, digital humanities, or related disciplines. Since TED talks are produced for a general audience, therefore, all the users were chosen not to be from an economics background as all the test videos are on the topic of economics. Admittedly participants are a rather cohesive group in the sense that they are all of an academic background. However, they are an adequate representation of the larger TED audience which is described as highly educated with an interest in scientific and intellectual pursuits [70,75].

7.5.2. Reviewers

A total of 7 reviewers participated in judging the summaries created by users. In a similar manner to the users; reviewers were also chosen not to be from an economics background and had postgraduate degrees in computer science, linguistics, or related disciplines.

7.6. Feedback Capturing, Annotations and Data For Analysis

7.6.1. Screen Capturing

User actions and their interaction with the representation was recorded via screen capturing and audio recording.

7.6.2. Summary Evaluation Task (Performed by Reviewers)

In synopsis writing task users produced a total of 24 summaries, 12 of them were created by using the RAAVE system while the rest were created using the baseline video player. Since 4 TED talks were used in the experiment there are 6 summaries created for each video.

There are two types of techniques to compare summaries, researchers often use automatic tools such as ROUGE [76]. The other technique is to use human evaluators. Bayomi et al. [77] observed that automatic techniques fall short in evaluating certain factors of summary quality compared to

humans. Therefore in order to evaluate the user produce synopses, the current experiment used a similar approach.

All 7 reviewers evaluated the summaries of all four TED talks. For each TED talk, they were provided with the video and the six summaries created of that TED talk. They were asked to do the following.

1. Watch the TED talk.
2. Evaluate each summary individually according to the characteristics listed in Table 5.
3. Rank the summaries in order of preferences (1 and 6).

Reviewers were asked to assign a rank between 1 to 6 to the synopsis of each video with the most preferred summary being 1st and the least preferred summary being 6th.

Table 5. Summary evaluation criteria.

Readability and Understandability: Whether the grammar and the spelling of the summary are correct and appropriate.
Extremely Bad - 1 2 3 4 5 - Excellent
Informativeness: How much information from the source video is preserved in the summary.
Extremely Bad - 1 2 3 4 5 - Excellent
Conciseness: As a summary presents a short text,
Extremely Bad - 1 2 3 4 5 - Excellent
Overall: The overall quality of the summary.
Extremely Bad - 1 2 3 4 5 - Excellent

7.6.3. Answers and Durations

For the answer search task following items were recorded for each user.

- The number of questions attempted.
- The number of questions correctly answered.
- Time taken to complete the task.
- Duration per correct answer (by normalization).

Normalizing duration per correct answer

To assess the efficiency of answer searching, measuring the average time to find an answer would not be appropriate since the videos are of different lengths. To normalize that, the percentage of video length is used. For example, if a user took 5 min to search for answers for a 20 min video, it is considered as 25% of the length of the video was required by the user to find all the answers. Dividing it by the number of correct answers gives the length of the video required per correct answer.

7.7. User Interactivity with the Two Systems

7.7.1. Baseline Player Logs

For the baseline video player, a log of user interactions with the video player was recorded using the SocialSkip system [78]. SocialSkip logs the standard interaction such as play, pause and seek etc. along with timestamps and other relevant meta-data on a server that can be downloaded as a CVS file.

7.7.2. Annotation of User Interactions

For the analysis, user interactions were annotated manually from video recordings. Table 6 shows an example for user# 12 interacting with the baseline system while performing the question search

task. The top row shows the minute of the experiment session. The left column shows the user's interaction with the system. In the example user is increasing the video play speed to 1.5×. The right column shows the user's action for the task at hand. In the example user spent the minute reading the questions and wrote the answer for a question.

Table 6. Annotation example for user interactions using baseline system.

Minute: 26-27	
Action System	**Action Task**
Play speed 1.5×. at 26:11	Reading questions
Ans. Q.13 at 26:21	

Table 7 shows the example for user# 12 interacting with RAAVE system. In the example user interacts with the word-cloud, summary, and transcript of segment no 4, 6 and 7. User also wrote the answer of Q.12 in that minute.

Table 7. Annotation example for user interaction using RAAVE system.

Minute: 12-13	
Action System	**Action Task**
Scrolling	
Seg 4 word cloud sci	
Seg 6 sum (sci)	
Seg 6 trans	
Seg 7 trns (sci)	
scolling	Ans. Q.12 at 12:03

8. Results

8.1. Results Hypothesis a (User Experience Assessment)

Hypothesis 1: Users have a better experience interacting with RAAVE compared to the baseline player.

Users were asked to fill the UEQ twice, once for RAAVE and once for the baseline system. The 26 questions of the UEQ can be categorized into following categories.

- Attractiveness
- Perspicuity
- Efficiency
- Dependability
- Stimulation
- Novelty

The one-sample Kolmogorov-Smirnov test showed that the user ratings provided for each category follows a normal distribution assumption with $p < 0.01$.

Figure 6 shows the comparison of user experience with both systems. Users scored RAAVE better in all categories except "Perspicuity" which is not surprising since the baseline system is much for familiar and simpler than RAAVE system. Figure 6 shows the T-Test score to assess if the difference between the two systems reported by users is statistically significant or not, as it can be seen that results are statistically significant in all but one category (Figure 6).

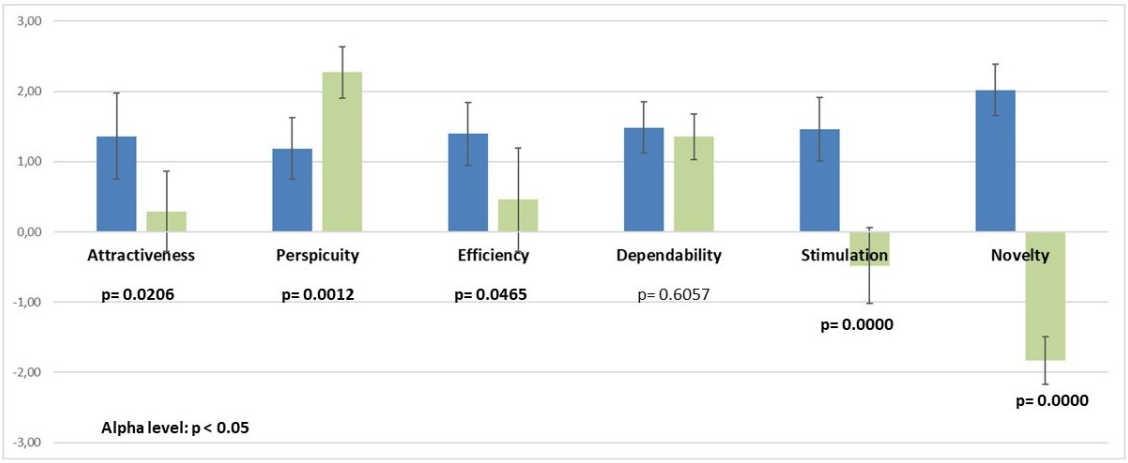

Figure 6. Blue (darker) bars represent RAAVE while green (lighter) bars represent the baseline system. Simple T-Test to check if the scale means of the two systems differ significantly.

8.2. Results Hypothesis 2 (Answer Search Task)

Hypothesis 2: RAAVE is better at allowing users to search for information in different parts of a video compared to baseline. By better, we mean the following:

- Users were able to answer the question more accurately with RAAVE compared to Baseline.
- Users were able to search the answers efficiently with RAAVE compared to Baseline.

Table 8, shows the results of answer search task. Table 8 shows the overall performance of users using both systems.

Overall results in Table 8 show that in terms of answering correctly users did better using the baseline systems 9.5 correct answers compared to 9.16 using RAAVE. However, for the two videos containing visual aid (ED+TP) users were able to perform better using RAAVE i.e., on average, the user performed better with RAAVE for difficult videos whereas their performance was better using the baseline system for easier videos. (detailed in Section 7.4 for discussion about easy and difficult videos).

In terms of efficiency, users seem to perform better using the RAAVE system, as the normalized duration per correct question is lower for RAAVE compared to Baseline. The duration per question is calculated by measuring the normalized duration (Section 7.6.3) spent by the user performing the task, divided by correct answers (lower is better).

Table 8. Answer search task results (overall performance).

Row#	Videos	RAAVE		Baseline	
		Correct	Per Question	Correct	Per Question
1	Overall Avg. (avg. corr. / avg. dur)	9.16	9.57	9.50	10.23
2	Average all users	9.17	10.04	9.50	10.75
3	Median all users	10	8.80	10	10.60
4	CF+PC (Avg.) (no slides)	10	8.12	11.16	8.46
5	ED+TP (Avg.) (with slides)	8.33	11.21	7.83	12.75

It can be seen in Table 8 that while on average user answer more correctly using the baseline system (9.50 using baseline vs. 9.17 using RAAVE) in terms of efficiency user spent less time per correct answer using RAAVE compared to the baseline (10.04 using RAAVE vs. 10.75 using Baseline (lower is better)). The Student T-Test p-value of the result is 0.27 which means the results are not statistically significant.

8.3. Results Hypothesis C (Synopsis Writing Task)

Hypothesis 3: Users can get a quicker understanding of the content of video using RAAVE compared to the baseline player.

Since the time allowed for users to consume the content was fixed at 20% of the video length. The comparison of user performance for synopsis writing task is done as:

- Comparison of reviewer ratings to the synopsis produced by using both systems.
- The likelihood of a synopsis produced by RAAVE be given top rank by reviewers is higher compared to the Baseline.

Table 9 shows the average score for each characteristic. The first row shows the overall scores i.e., the average of scores against all the summaries of 4 videos. The next four rows show the average score of the summaries of an individual video. The last two rows combine the average score of videos with and without visual aid. "CF+PC" shows the average score of [72,73] since these two videos do not contain any visual aid and 'ED+TP' shows the average score of [69,71].

In addition to the rating of the user-produced summaries individually, reviewers were also asked to rank the summaries in order of their preference for 'ranked data analysis'. Users produced 6 summaries per video. For each video, reviewers assigned a rank between 1 and 6 to produced summaries (1 for most preferred and 6 to the least preferred).

The ranked data analysis is performed using Mallows–Bradley Terry (MBT) method [79]. The MBT results in Table 9 follow the same pattern as in the question search task. The scores (θ_n where $n = 6$ as there are six summaries ranked by a reviewer. $\Theta_{MBT} = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$ is the estimated score for each summary. A summary with a higher score is better than a summary with a lower score) are slightly better with the Baseline system overall. However, better results are achieved using RAAVE system with videos that contain visual aid and are relatively more technical and difficult to comprehend than the other two.

Table 10 shows the results for each video. Est. column lists the estimated likelihood for the summary produced to be ranked no.1 by reviewers. For two of the videos CF and ED the RAAVE has a higher probability to produce the top summary while for the other two Baseline scored higher.

Table 11 shows the overall likelihood for each system instead of individual summaries. It can be seen that the RAAVE has scored higher for both kinds of videos i.e., the likelihood that the top rank will be assigned to a summary produced using the RAAVE system by reviewers.

Table 9. Average scores for each characteristic.

	Readability and Understandability		**Informativeness**		**Conciseness**		**Overall**	
	RAAVE	**Baseline**	**RAAVE**	**Baseline**	**RAAVE**	**Baseline**	**RAAVE**	**Baseline**
All	3.50	3.64	3.01	3.12	3.50	3.36	3.19	3.20
CF	3.32	3.43	2.61	3.21	3.43	3.71	3.04	3.00
ED	3.65	4.00	3.15	2.93	3.46	3.07	3.23	3.13
PC	3.64	3.89	2.86	3.00	3.50	3.36	3.00	3.21
TP	3.42	3.25	3.83	3.33	3.75	3.33	3.67	3.33
CF+PC	3.43	3.74	2.69	3.07	3.45	3.48	3.02	3.14
ED+TP	3.58	3.54	3.37	3.18	3.55	3.23	3.37	3.26

Table 10. Θ_{MBT} parameters for each summary, indicating a summary rank (calculated using MBT) among other summaries. The top two θ are in bold.

TP			CF			ED			PC		
Θ_{MBT}	User	Sys	Est.	User	Sys.	Θ_{MBT}	User	Sys.	Θ_{MBT}	User	Sys.
0.1076	1	B	0.125	1	R	0.1891	2	B	0.08249	2	R
0.2314	3	R	0.1647	3	B	**0.232**	5	R	**0.50695**	4	B
0.164	6	B	0.1125	4	R	**0.1984**	6	R	0.10486	5	B
0.0382	7	B	**0.2443**	7	R	0.1574	8	B	**0.13291**	8	R
0.1873	9	R	**0.1887**	9	B	0.1179	11	R	0.04768	10	B
0.2716	12	B	0.1647	10	R	0.1052	12	R	0.12511	11	B

Table 11. Probability that reviewers are likely to rank this as no.1. The best results are in bold for each video type.

Videos	RAAVE	Baseline
TP+ED (slides)	**0.54**	0.46
CF+PC (no slides)	**0.57**	0.43

9. Discussion

As explained in section Section 6, the user study is designed to stress test the prototype system thereby the proposed template-based approach, still the results of the experiment are encouraging.

For efficiently searching for information within video content, users spent less time per correct answer compared to the baseline system (Table 8). Overall users were able to answer more correctly using the baseline system, it is because due to the design of the experiment it was easier for them to watch the whole video and answer the questions in parallel. RAAVE got better results for difficult videos compared to the easy ones (Table 8). In terms of search strategies user who used a mixture of query box and content interaction were able to answer more accurately while spending less time on the task (row 3 and 4 of Table). This can be used in further streamlining the representation of the video i.e., using templates that encourage users to interact more with content.

Regarding quickly getting a better understanding of the essence of the video. The results are moderately encouraging for RAAVE system. The likelihood that the synopsis creates using RAAVE would be ranked 1st by reviewers was higher compared to the baseline player even though the margin is not very wide (Table 11). In terms of quality criterion overall synopsis produced using the baseline system were scored higher compared to RAAVE. However, as it was the case in answer search task RAAVE scored better for difficult videos (TP+ED) on all the 4, quality criterion (last row of Table 9). In terms of interaction strategies, the 3 users who only consumed the automatically generated summaries on average scored better than others.

Hence the proposed approach provides advantages in terms of providing a flexible and engaging experience to the user during exploration tasks and provides advantages in terms of spending less time searching for information and have a better understanding of video by choosing both the modality and amount of detail to consume the content. While it was initially assumed that giving users the ability to consume different modalities in parallel would be beneficial, e.g., it is possible for the user to listen to one segment while reading the summary of another. However, the results suggest that such a strategy does not always yield optimal performance.

In terms of user experience with the system, despite the lack of familiarity and other limitations of RAAVE users had a productive experience with the RAAVE system. Users rated the RAAVE system more favorably by a wide margin except in the category of dependability although RAAVE's score is still higher than Baseline player. It is not surprising as due to its familiarity, simple nature, and wide availability, the regular video player is very dependable i.e., it does the simple things it does quite

well, whereas RAAVE provided a lot of options and the UI was not fully matured. With a better UI and more practice, the user exploration experience with RAAVE is bound to get better.

While RAAVE is designed to represent the content of a video in contrast to video search and retrieval systems such as the ones in Video Browser Showdown or VBS [12]. The presented study has shown the potential of the developed prototype and by extension RAAVE approach to be used in conjunction or as a supplementary system to any video retrieval and searching system.

10. Conclusions and Future Directions

This article proposed an approach to transform a video into an interactive multimedia document. Transforming a video into an interactive document opens up new ways to explore the content of the video as users in addition to watching the video, can consume it in a combination of different modalities and amount of detail, better suited to the context.

This article has presented the design and evaluation of RAAVE which transforms a video into an interactive multimodal document to open up new ways to explore its content. The experiments performed in this study currently are narrowed down in the following ways:

- Only presentation style TED videos are used.
- Only query-based scenario for exploration was evaluated.

There are many possible directions to pursue further research work. Future directions can be divided into two broad categories:

- Enhancing feature extractions.
- Applying RAAVE in different scenarios.

in terms of feature extraction enhancement, the future plan is to incorporate more multimodal features in the representation and evaluate the potential enhancement in the exploration. Some examples are:

- Visual features such as facial expressions, body movements.
- Audio/paralinguistic features.
- Linguistic features such as semantic uplift of topic concepts etc.

In terms of applying RAAVE to different scenarios, it is our intention to apply the proposed approach on a variety of video content e.g.,

- Massive Online Open Courses (MOOC) videos.
- Training videos. (Lynda, misc. corporate training videos).
- Instructional tutorials (how to fix a bike, how to apply makeup etc.).
- News footage and documentaries.
- Lifelogging videos (meeting recordings, conference calls, video chats)

As an example consider lifelogging videos particularly meeting recording or recordings of conference calls or video chats. The proposed template-based multimodal representation can be used to provide the ability to explore the content of meeting in a nonlinear and multimodal manner. Girgensohn et al. proposed a hypervideo-based approach to explore meeting recordings [42]. As detailed in state of the art (Section 2), RAAVE extends the idea of nonlinear exploration of video by transforming it into the multimedia document. Similarly, for meeting recordings, the template approach can be applied to create a multimedia brief based on not only the topic of interest or speaker choice [22,42] but also choose the amount of detail and choice of modality by choosing an appropriate template based on user preference or end user device.

In addition to a variety of content, the other dimension is the application of the approach in exploration task scenarios. In future, we intend to test the approach in a variety of situations

e.g., automatic curation of news article or multimedia essay from video footage based on a personalization model instead of waiting for the user to execute a query.

Another interesting use of the proposed approach is to transform the video content for professional use cases. An example could be allowing the ability to search for information within long video footage and automatically or semi-automatically curating a new multimedia artefact which may be a video, or it may be a multimedia document.

Finally, we hope that the proposed idea of transforming content based on context can be expanded to content other than video. The design factors of the representation engine can be used to search heterogeneous data-sources, which could be structured or semi-structured, where the information extracted can be automatically represented or curated as an interactive multimedia document and help create digital narratives on demand.

Author Contributions: Conceptualization, F.A.S. and O.C.; formal analysis, F.A.S. and F.H.; funding acquisition, O.C.; methodology, F.A.S. and F.H.; software, F.A.S.; supervision, O.C.; validation, S.L. and O.C.; writing—original draft, F.A.S. and F.H.; writing—review and editing, F.A.S., F.H., S.L. and O.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661, SAAM project at the University of Edinburgh, UK, and "ADAPT 13/RC/2106" project (http://www.adaptcentre.ie/) at Trinity College Dublin, the University of Dublin, Ireland.

Acknowledgments: The authors would like to acknowledge all our colleagues who participated in this user study and funding bodies.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. User Experience Questionnaire

Please make your evaluation now.

For the assessment of the product, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:

attractive	○	⊗	○	○	○	○	○	unattractive

This response would mean that you rate the application as more attractive than unattractive.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Figure A1. *Cont*.

Please assess the product now by ticking one circle per line.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Figure A1. Full list of User Experience Questionnaire Questions.

Appendix A.2. Questions for the Video 'New Thoughts on Capital in the Twenty-First Century'

MCQ for Ted Talk Titled: "New thoughts on capital in the twenty-first century" by Thomas Piketty

The Video has been segmented into 10 smaller segments. Following are some questions which can be answer with content presented in the video. Please read the questions write down the segment in which you think you are most likely to find the answer to the question.

1. In what segment(s) does the author reiterates the recurring theme of this talk.

   Segment # _____

2. The factors in income inequality being higher in US compared to Europe.

   Segment # _____

3. What made the swiss show flexibility in bank secrecy?

   Segment # _____

4. Which one is a criticism on the presenter thesis

   Segment # _____

5. What is the Data source used by the presenter?

   Segment # _____

6. Some economists argue in support of inequality that it's an engine of capitalism

   Segment # _____

7. The least efficient way of decreasing inequality is starting wars

   Segment # _____

8. The growth rate of economy has been unusually high in certain countries during

Figure A2. *Cont.*

Segment # _____

9. The World wars and their aftermath has Decreased inequality

   Segment # _____

10. Decrease in Capital Gains has caused an increase in economic growth

    Segment # _____

11. In Pre-Industrial society the growth rate of economy was traditionally close to zero.

    Segment # _____

12. In 21st century the Top 10% population has the following share of global income.

    Segment # _____

13. Jane Austen is mentioned in.

    Segment # _____

14. In 21st century the Top 10% population has the following share of global wealth.

Segment # _____

Figure A2. Full list of Questions for Thomas Piketty (TP) video for the answer search task.

References

1. Hong, R.; Tang, J.; Tan, H.K.; Ngo, C.W.; Yan, S.; Chua, T.S. Beyond search Event Driven summarization of web videos. *ACM Trans. Multimed. Comput. Commun. Appl.* **2011**, *7*, 1–18. [CrossRef]
2. Mujacic, S.; Debevc, M.; Kosec, P.; Bloice, M.; Holzinger, A. Modeling, design, development and evaluation of a hypervideo presentation for digital systems teaching and learning. *Multimed. Tools Appl.* **2012**, *58*, 435–452. [CrossRef]
3. Masneri, S.; Schreer, O. SVM-based Video Segmentation and Annotation of Lectures and Conferences. In Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, 5–8 January 2014; IEEE: Lison, French, 2014; pp. 425–432.
4. Darrell Etherington People Now Watch 1 Billion Hours of YouTube Per Day. Available online: https://techcrunch.com/2017/02/28/people-now-watch-1-billion-hours-of-youtube-per-day/ (accessed on 10 April 2020)
5. CISCO. The Zettabyte Era: Trends and Analysis. *Cisco* **2017**, 1–29. [CrossRef]
6. Shen, J.; Cheng, Z. Personalized video similarity measure. *Multimed. Syst.* **2010**, *17*, 421–433. [CrossRef]
7. Halvey, M.; Vallet, D.; Hannah, D.; Jose, J.M. Supporting exploratory video retrieval tasks with grouping and recommendation. *Inf. Process. Manag.* **2014**, *50*, 876–898. [CrossRef]

8. Schoeffmann, K.; Hudelist, M.A. Video Interaction Tools: A Survey of Recent Work. *ACM Comput. Surv.* **2015**, *48*. [CrossRef]
9. Petan, A.S.; Petan, L.; Vasiu, R. Interactive Video in Knowledge Management: Implications for Organizational Leadership. *Procedia- Soc. Behav. Sci.* **2014**, *124*, 478–485. [CrossRef]
10. Ericsson. *TV AND MEDIA 2016, An Ericsson Consumer and Industry Insight Report*; Technical Report November; Erricsson: Stockholm, Sweden, 2016.
11. Waitelonis, J.; Sack, H. Towards exploratory video search using linked data. *Multimed. Tools Appl.* **2012**, *59*, 645–672. [CrossRef]
12. Cobârzan, C.; Schoeffmann, K.; Bailer, W.; Hürst, W.; Blažek, A.; Lokoč, J.; Vrochidis, S.; Barthel, K.U.; Rossetto, L. Interactive video search tools: A detailed analysis of the video browser showdown 2015. *Multimed. Tools Appl.* **2017**. [CrossRef] [PubMed]
13. O'Brien, H.L.; Toms, E.G. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Inf. Process. Manag.* **2013**, *49*, 1092–1107. [CrossRef]
14. Dobrian, F.; Awan, A.; Joseph, D.; Ganjam, A.; Zhan, J.; Berkeley, U.C. Understanding the Impact of Video Quality on User Engagement. *World* **2011**, 362–373. [CrossRef]
15. Guo, P.J.; Kim, J.; Rubin, R. How Video Production Affects Student Engagement : An Empirical Study of MOOC Videos. In Proceedings of the 1st ACM Conference on Learning at Scale (L@S 2014), Atlanta, GA, USA, 4–5 March 2014; doi:10.1145/2556325.2566239. [CrossRef]
16. Benini, S.; Migliorati, P.; Leonardi, R. Statistical Skimming of Feature Films. *Int. J. Digit. Multimed. Broadcast.* **2010**, *2010*, 1–11. [CrossRef]
17. Haesen, M.; Meskens, J.; Luyten, K.; Coninx, K.; Becker, J.H.; Tuytelaars, T.; Poulisse, G.J.; Pham, P.T.; Moens, M.F. Finding a needle in a haystack: An interactive video archive explorer for professional video searchers. *Multimed. Tools Appl.* **2011**, *63*, 331–356. [CrossRef]
18. Laugwitz, B.; Held, T.; Schrepp, M. Construction and Evaluation of a User Experience Questionnaire. *Hci Usability Educ. Work.* **2008**, 63–76._6. [CrossRef]
19. Ruotsalo, T.; Jacucci, G.; Myllymäki, P.; Kaski, S. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* **2015**, *58*, 86–92. [CrossRef]
20. Zhang, D.; Nunamaker, J.F. A natural language approach to content-based video indexing and retrieval for interactive e-Learning. *IEEE Trans. Multimed.* **2004**, *6*, 450–458. [CrossRef]
21. Luz, S.; Roy, D.M. Meeting browser: A system for visualising and accessing audio in multicast meetings. In Proceedings of the 1999 IEEE Third Workshop on Multimedia Signal Processing, Copenhagen, Denmark, 13–15 September 1999; pp. 587–592. [CrossRef]
22. Luz, S.; Masoodian, M. A Model for Meeting Content Storage and Retrieval. In Proceedings of the 11th International Conference on Multi-Media Modeling (MMM 2005), Melbourne, Australia, 12–14 January 2005; Chen, Y.P.P., Ed.; IEEE Computer Society: Melbourne, Australia, 2005; pp. 392–398. [CrossRef]
23. Barthel, R.; Ainsworth, S.; Sharples, M. Collaborative knowledge building with shared video representations. *Int. J. Hum. Comput. Stud.* **2013**, *71*, 59–75. [CrossRef]
24. Merkt, M.; Schwan, S. Training the use of interactive videos: Effects on mastering different tasks. *Instr. Sci.* **2014**, *42*, 421–441. [CrossRef]
25. Pavel, A.; Reed, C.; Hartmann, B.; Agrawala, M. Video digests: A browsable, skimmable format for informational lecture videos. In Proceedings of the Symposium on User Interface Software and Technology, Honolulu, HI, USA, 5–8 October 2014; pp. 573–582. [CrossRef]
26. Meixner, B.; Gold, M. Second-Layer Navigation in Mobile Hypervideo for Medical Training. In Proceedings of the MultiMedia Modeling: 22nd International Conference (MMM 2016), Miami, FL, USA, 4–6 January 2016; Volume 9516, pp. 382–394. [CrossRef]
27. Sauli, F.; Cattaneo, A.; van der Meij, H. Hypervideo for educational purposes: A literature review on a multifaceted technological tool. *Technol. Pedagog. Educ.* **2017**, *5139*, 1–20. [CrossRef]
28. Nelson, T.H. Complex information processing. In Proceedings of the 1965 20th national conference, New York, NY, USA, 24–26 August 1965; pp. 84–100. [CrossRef]
29. Boissiere, G. Automatic Creation of Hypervideo News Libraries for the World Wide Web. In Proceedings of the HYPERTEXT '98 Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, USA, 20–24 June 1998; pp. 279–280.

30. Finke, M.; Balfanz, D. A reference architecture supporting hypervideo content for ITV and the internet domain. *Comput. Graph. (Pergamon)* **2004**, *28*, 179–191. [CrossRef]
31. Stahl, E.; Finke, M.; Zahn, C. Knowledge Acquisition by Hypervideo Design : An Instructional Program for University Courses. *J. Educ. Multimed. Hypermed.* **2006**, *15*, 285–302.
32. Hoffmann, P.; Herczeg, M. Hypervideo vs. Storytelling: Integrating Narrative Intelligence into Hypervideo. In Proceedings of the Third International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE 2006), Darmstadt, Germany, 4–6 December 2006; pp. 37–48._4. [CrossRef]
33. Aubert, O.; Champin, P.A.; Prié, Y.; Richard, B. Canonical processes in active reading and hypervideo production. *Multimed. Syst.* **2008**, *14*, 427–433. [CrossRef]
34. Hildebrand, M.; Hardman, L. Using Explicit Discourse Rules to Guide Video Enrichment. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 461–464.
35. Leggett, M.; Bilda, Z. Exploring design options for interactive video with the Mnemovie hypervideo system. *Des. Stud.* **2008**, *29*, 587–602. [CrossRef]
36. Shipman, F.; Girgensohn, A.; Wilcox, L. Authoring, viewing, and generating hypervideo. *ACM Trans. Multimed. Comput. Commun. Appl.* **2008**, *5*, 1–19. [CrossRef]
37. Tiellet, C.A.; Pereira, A.G.; Reategui, E.B.; Lima, J.V.; Chambel, T. Design and evaluation of a hypervideo environment to support veterinary surgery learning. In Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT '10), Toronto, ON, Canada, 13–16 June 2010; pp. 213–222. [CrossRef]
38. Sadallah, M.; Aubert, O.; Prié, Y. CHM: An annotation- and component-based hypervideo model for the Web. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 869–903. [CrossRef]
39. Bulterman, D.; Rutledge, L. *SMIL 3.0*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009; p. 507.
40. Neto, C.d.S.S.; Soares, L.F.G. Reuse and imports in Nested Context Language. In Proceedings of the XV Brazilian Symposium on Multimedia and the Web (WebMedia '09), Fortaleza, Brazil, 5–7 October 2009; pp. 1–8. [CrossRef]
41. Meixner, B.; Matusik, K.; Grill, C.; Kosch, H. Towards an easy to use authoring tool for interactive non-linear video. *Multimed. Tools Appl.* **2014**, *70*, 1251–1276. [CrossRef]
42. Girgensohn, A.; Marlow, J.; Shipman, F.; Wilcox, L. Guiding Users through Asynchronous Meeting Content with Hypervideo Playback Plans. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (HT '16), Halifax, NS, Canada, 10–13 July 2016; pp. 49–59. [CrossRef]
43. Leiva, L.A.; Vivó, R. Web browsing behavior analysis and interactive hypervideo. *ACM Trans. Web* **2013**, *7*, 1–28. [CrossRef]
44. Meixner, B. Hypervideos and Interactive Multimedia Presentations. *ACM Comput. Surv.* **2017**, *50*, 1–34. [CrossRef]
45. Guan, G.; Wang, Z.; Mei, S.; Ott, M.A.X.; He, M.; Feng, D.D. A Top-Down Approach for Video Summarization. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2014**, *11*, 1–21. [CrossRef]
46. de Avila, S.E.F.; Lopes, A.P.B.; da Luz, A.; de Albuquerque Araújo, A. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.* **2011**, *32*, 56–68. [CrossRef]
47. Almeida, J.; Leite, N.J.; Torres, R.D.S. Online video summarization on compressed domain. *J. Vis. Commun. Image Represent.* **2013**, *24*, 729–738. [CrossRef]
48. Zhang, Y.; Zhang, L.; Zimmermann, R. Aesthetics-Guided Summarization from Multiple User Generated Videos. *ACM Trans. Multimed. Comput. Commun. Appl.* **2015**, *11*, 1–23. [CrossRef]
49. Belo, L.d.S.; Caetano, C.A.; do Patrocínio, Z.K.G.; Guimarães, S.J.F. Summarizing video sequence using a graph-based hierarchical approach. *Neurocomputing* **2016**, *173*, 1001–1016. [CrossRef]
50. Chen, F.; Vleeschouwer, C.D.; Cavallaro, A. Resource Allocation for Personalized Video Summarization. **2014**, *16*, 455–469. [CrossRef]
51. Kim, D.J.; Frigui, H.; Fadeev, A. A generic approach to semantic video indexing using adaptive fusion of multimodal classifiers. *Int. J. Imaging Syst. Technol.* **2008**, *18*, 124–136. [CrossRef]
52. Wang, J.; Duan, L.; Liu, Q.; Lu, H.; Jin, J.S. A multimodal scheme for program segmentation and representation in broadcast video streams. *IEEE Trans. Multimed.* **2008**, *10*, 393–408. [CrossRef]
53. Hosseini, M.S.; Eftekhari-Moghadam, A.M. Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. *Appl. Soft Comput.* **2013**, *13*, 846–866. [CrossRef]

54. Evangelopoulos, G.; Zlatintsi, A.; Potamianos, A.; Maragos, P.; Rapantzikos, K.; Skoumas, G.; Avrithis, Y. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimed.* **2013**, *15*, 1553–1568. [CrossRef]
55. Jin, Q.; Chen, J.; Chen, S.; Xiong, Y.; Hauptmann, A. Describing videos using multi-modal fusion. In Proceedings of the 2016 ACM Multimedia Conference (MM 2016), Amsterdam, The Netherlands, 15–19 October 2016; pp. 1087–1091. [CrossRef]
56. Liu, C.; Mao, J.; Sha, F.; Yuille, A. Attention correctness in neural image captioning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017), San Francisco, CA, USA, 4–9 February 2017; pp. 4176–4182. Available online: http://xxx.lanl.gov/abs/1605.09553 (accessed on 4 February 2019).
57. Yadav, K.; Shrivastava, K.; Mohana Prasad, S.; Arsikere, H.; Patil, S.; Kumar, R.; Deshmukh, O. Content-driven Multi-modal Techniques for Non-linear Video Navigation. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15), Marina del Ray, CA, USA, 17–20 March 2019; pp. 333–344. [CrossRef]
58. Hudelist, M.; Schoeffmann, K.; Xu, Q. Improving interactive known-item search in video with the keyframe navigation tree. *MultiMedia Model.* **2015**, *8935*, 306–317._27. [CrossRef]
59. Salim, F.; Haider, F.; Conlan, O.; Luz, S. *An Alternative Approach to exploring a Video*; LNAI; Springer: Cham, Sweden, 2017; Volume 10458._10. [CrossRef]
60. Monserrat, T.; Zhao, S.; McGee, K.; Pandey, A. NoteVideo: Facilitating navigation of blackboard-style lecture videos. In Proceedings of the CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 1139–1148. [CrossRef]
61. Choi, F. Advances in Domain Independent Linear Text Segmentation. In Proceedings of the NAACL 2000, Seattle, WA, USA, 29 April–4 May 2000; Association for Computational Linguistics: Seattle, WA, USA, 2000; pp. 26–33.
62. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MA, USA, 23–24 June 2014; pp. 55–60. [CrossRef]
63. Bradski, G. Dr. Dobb's Journal, San Francisco. *The OpenCV Library*; 2000.
64. Lienhart, R.; Maydt, J. An extended set of Haar-like features for rapid object detection. In Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22–25 September 2002; Volume 1. [CrossRef]
65. Autosummarizer. Available online: autosummarizer.com (accessed on 23 April 2020).
66. Steinbock, D. Available online: http://tagcrowd.com/ (accessed on 23 April 2020).
67. Velasco, R. *Apache Solr: For Starters*; CreateSpace Independent Publishing Platform: North Charleston, SC, USA, 2016..
68. McCandless, M.; Hatcher, E.; Gospodnetic, O. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*; Manning Publications Co.: Greenwich, CT, USA, 2010.
69. Piketty, T. New Thoughts on Capital in the Twenty-First Century. 2014. www.Ted.com (accessed on 23 April 2020).
70. Scotto di Carlo, G. The role of proximity in online popularizations: The case of TED talks. *Discourse Stud.* **2014**, *16*, 591–606. [CrossRef]
71. Duflo, E. Social Experiments to Fight Poverty. 2010. www.Ted.com (accessed on 23 April 2020).
72. Collier, P. The Bottom Billion. 2008. www.Ted.com (accessed on 23 April 2020).
73. Freeland, C. The Rise of the New Global Super-Rich. 2013. www.Ted.com (accessed on 23 April 2020).
74. Gravier, G.; Ragot, M.; Laurent, A.; Bois, R.; Jadi, G.; Jamet, E.; Monceaux, L. Shaping-Up Multimedia Analytics: Needs and Expectations of Media Professionals. In Proceedings of the International Conference on Multimedia Modeling, Miami, GL, USA, 4–6 January 2016; Volume 9516; pp. 303–314. [CrossRef]
75. Sugimoto, C.R.; Thelwall, M.; Larivière, V.; Tsou, A.; Mongeon, P.; Macaluso, B. Scientists Popularizing Science: Characteristics and Impact of TED Talk Presenters. *PLoS ONE* **2013**, *8*. [CrossRef] [PubMed]
76. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the workshop on text summarization branches out (WAS 2004), Barcelona, Spain, 25–26 July 2004; pp. 25–26.

77. Bayomi, M.; Levacher, K.; Ghorab, M.R. Text Summarization and Speech Synthesis for the Automated Generation of Personalized Audio Presentations. In *International Conference on Applications of Natural Language to Information Systems*; Springer: Cham, Sweden, 2016; Volume 9612, pp. 187–199. [CrossRef]
78. Chorianopoulos, K. SocialSkip: Pragmatic understanding within web video. In Proceedings of the 9th European Conference on Interactive TV and Video, Lisbon, Portugal, 29 June 2011; pp. 25–28. [CrossRef]
79. Critchlow, D.E.; Fligner, M.A. Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika* **1991**, *56*, 517–533. [CrossRef]